Outils pour lexicographes : application à la lexicographie explicative et combinatoire

Gilles Sérasset, Alain Polguère

▶ **To cite this version:**

Gilles Sérasset, Alain Polguère. Outils pour lexicographes : application à la lexicographie explicative et combinatoire. RIAO'97, 1997, Montréal, Canada. pp.701-708. hal-00966344

HAL Id: hal-00966344
https://hal.science/hal-00966344

Submitted on 26 Mar 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Outils pour lexicographes : application à la lexicographie explicative et combinatoire

Gilles Sérasset

GETA-CLIPS-IMAG (UJF & CNRS)
BP 53
38041 Grenoble Cedex 9

Tél. : 04.76.51.43.80 - Fax : 04.76.51.44.05
Courriel : Gilles.Serasset@imag.fr

Alain Polguère

Département de linguistique et traduction
Université de Montréal
CP 6128, succ. Centre Ville
H3C 3J7 Montréal Québec Canada
Tél. : (514) 343 6111 - Fax : (514) 343 2284
Courriel : polguera@ere.umontreal.ca

Résumé

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmente avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires.

Nous nous sommes donc intéressé à la construction d'outils pour lexicographes et lexicologues. Afin d'avoir une bonne compréhension des problèmes qui se posent, nous avons décidé d'informatiser un dictionnaire complexe, contenant de nombreuses informations structurées, le dictionnaire explicatif et combinatoire du français contemporain (DEC). Le DEC étant un travail de lexicologie, il ne s'agit donc pas à proprement parler d'un dictionnaire, mais plutôt d'un ensemble d'entrées destinées à illustrer une théorie linguistique. Ce ne sont donc pas les données que l'on va informatiser, mais le processus de rédaction de ces données.

Cette action a été menée en collaboration entre le GETA-CLIPS (université Joseph Fourier – Grenoble 1) et le GRESLET (université de Montréal), grâce aux soutiens du réseau LTT de l'AUPELF-UREF et des ministères français et canadiens des affaires étrangères (coopérations franco-québecoises en ingénierie linguistique).

Mots clés

Traitement Automatique des Langues Naturelles ; base lexicale multilingue ; lexicographie ; dictionnairique.

1. Introduction

Dans le domaine de l'ingénierie linguistique et de la connaissance, le problème des ressources lexicales et linguistiques s'est toujours posé. Néanmoins, l'avancée des techniques du Traitement Automatique des Langues Naturelles (TALN) l'a rendu plus sensible. Il nous faut maintenant pouvoir répondre à des besoins importants en terme de quantité, de qualité et de complexité. La complexité et la diversité des informations requises augmente avec les exigences des outils de TALN ainsi qu'avec le développement de nouvelles applications (humaines ou machinales). Si la récupération (semi)automatique d'information lexicale est une piste, elle ne pourra remplacer la création manuelle de dictionnaires.

Nous nous sommes donc intéressé à la construction d'outils pour lexicographes et lexicologues. Afin d'avoir une bonne compréhension des problèmes qui se posent, nous avons décidé d'informatiser un dictionnaire complexe, contenant de nombreuses informations structurées, le dictionnaire explicatif et combinatoire du français contemporain (DEC). Le DEC étant un travail de lexicologie, il ne s'agit donc pas à proprement parler d'un dictionnaire, mais plutôt d'un ensemble d'entrées destinées à illustrer une théorie linguistique. Ce ne sont donc pas les données que l'on va informatiser, mais le processus de rédaction de ces données.

Cette action a été menée en collaboration entre le GETA-CLIPS (université Joseph Fourier – Grenoble 1) et le GRESLET (université de Montréal), grâce aux soutiens du réseau LTT de l'AUPELF-UREF et des ministères français et canadiens des affaires étrangères (coopérations franco-québecoises en ingénierie linguistique).

Nous présentons les outils et méthodes que nous avons adopté pour la construction de dictionnaires. Nous montrerons ensuite en détail l'outil DECID, un éditeur spécialisé pour le dictionnaire explicatif et combinatoire, défini par [Mel'ãuk et al. 95]. Nous aborderons ce faisant la stratégie que nous avons adopté pour l'informatisation du dictionnaire explicatif et combinatoire.

2. Outils pour la construction de lexiques

2.1. Généralités

Dans le processus de construction d'un lexique, différents intervenants sont concernés :

- le **lexicologue** défini les informations qui seront contenues dans le lexique, spécifie leurs forme et donne les critères permettant de définir les unités du lexique.
- l'**informaticien** construit les outils spécifiques au lexique ainsi défini et met au point la méthodologie qui sera utilisée lors de la construction du lexique. Il construit de plus les interfaces nécessaire au lexicographe.
- le **lexicographe** construit le lexique selon les spécifications ainsi faites. Il va construire les unités du lexique et/ou compléter des unités déjà existantes.

La constructions d'outils pour le lexique pose des problèmes informatiques forts du fait de la masse des informations à construire. La construction d'un dictionnaire est un travail mené en collaboration par différents lexicographes qui doivent respecter une cohérence, non seulement pour la forme spécifiée par le lexicologue (abréviations, balises…), mais aussi sur le fond (même critère de sélection des sens, mêmes critères de décomposition en entrées et sous entrées dans le cas d'homographes…). Enfin, les choix faits par certains lexicographes peuvent influer sur les décisions que devront prendre d'autres lexicographes (liens syntaxiques ou sémantiques entre entrées).

Les outils informatiques construits doivent tenir compte de l'aspect *distribué* du travail de lexicographie.

Pour construire une entrée de dictionnaire, le lexicographe doit rechercher les différents usages de l'entrée donnée. Il ne peut se contenter de son "intuition" linguistique. Cette validation passe par une recherche d'informations différentes dans des sources différentes (corpus, dictionnaires existants, locuteur natif…). À l'heure actuelle, ces sources sont souvent dispersées sous différentes plates-formes logicielles et matérielles (lorsqu'elles sont disponibles sous forme informatique). Un outil idéal de lexicographie doit donc pouvoir *intégrer* les différentes sources d'information brute sur la plate-forme d'édition.

Enfin, lors du travail de lexicographie, il peut arriver que le lexicologue souhaite modifier la structure du dictionnaire afin de mieux prendre en compte certains phénomènes qui ont été mal évalués ou sous-estimés. Cela peut se traduire par un ajout d'information, un retrait d'information, ou la modification de valeurs possible. Cette évolution de la structure du dictionnaire oblige à une reprise de tout ou partie des entrées déjà indexées. L'informaticien doit pouvoir fournir des outils permettant d'automatiser (au moins partiellement) ce travail de révision. Enfin, la modification de la structure du dictionnaire peut se traduire par un changement des interfaces d'édition dont dispose les lexicographe et par une modification des éventuels outils de vérifications automatiques de cohérence. Un outil pour lexicographe doit donc être suffisamment *paramétrable* et évolutif pour autoriser de tels changements.

2.2. Utilisation d'un éditeur général

Dans [Sérasset 94], nous définissons un système de gestion de bases lexicales multilingues nommé SUBLIM. Le système SUBLIM est destiné aux lexicologues qui peuvent, par l'intermédiaire de deux langages spécialisés, définir l'architecture lexicale (dictionnaires interlingues, monolingues, bilingues) et l'architecture linguistique (description des entrées de chacun des dictionnaires) de leur base.

L'architecture linguistique des entrées de chaque dictionnaire peut-être définie en utilisant des constructeurs de bases :

- structures de traits,
- graphes,
- arbres,
- automates,
- tables…

Les langages utilisés permettent de spécifier des structures complexes, comme celle du dictionnaire explicatif et combinatoire du français contemporain, en permettant de rester proche de la théorie linguistique sous-jacente. Nous essayons d'imposer le moins possible de contraintes informatique au lexicologue.

Le système SUBLIM propose des outils pour définir et gérer une base lexicale (vérification de cohérence, récupération de données, export de données…), mais il ne propose pas encore d'interface pour les lexicographes. Nous avons néanmoins fait des expériences dans ce domaine.

Les structures de la plupart des dictionnaires que nous avons été amenés à étudier étaient relativement simples et ne justifiaient pas la création d'un outils spécialisé. Nous avons donc préféré utiliser un traitement de texte répandu et disponible sur différentes plates-formes : Microsoft Word™. Chaque unité d'information est données sous forme de paragraphe dans un style particulier (figure 1).

Il nous est possible, à partir d'une base lexicale définie avec SUBLIM, de créer un modèle de document qui contient les styles correspondant aux unités d'information des entrées, ainsi que de nombreuses macro assistant le lexicographe dans l'édition des entrées. Grâce à la description en dictionnaire, on peut proposer des fenêtres permettant la sélection de catégories (ce qui évite les erreurs dans les abréviation). De plus, il nous est possible de calculer l'ensemble des styles valides à la suite d'un élément d'information. ainsi, dans l'exemple ci-dessous, le style de prononciation sera automatiquement inséré à la suite d'une entrée.

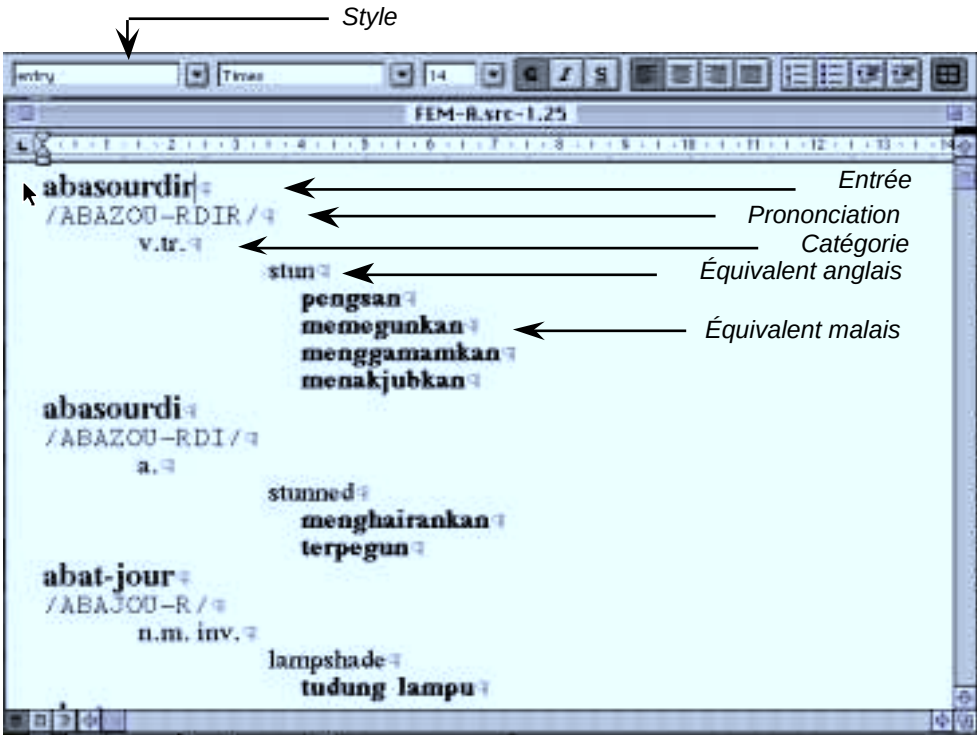

Figure 1 : Exemple de fichier d'édition d'un dictionnaire Français Anglais Malais.

À partir des informations présentes dans la base, on peut ensuite créer un fichier de travail (grâce à un export en format RTF) et le donner à un lexicographe. Une fois son travail accompli, nous le récupérons sous forme RTF et l'importons automatiquement dans la base lexicale.

Cette architecture n'utilise, sur le poste de travail du lexicographe, que des outils standards (un traitement de texte très répandu) et présente de nombreux avantages :

- Le lexicographe peut travailler sous MacOS ou sous Windows sans que nous ayons eu à programmer d'outil portable.
- Il dispose d'une vue globale de l'extrait de dictionnaire avec lequel il travaille. Il peut corriger très rapidement les erreurs qu'il détecte et peut s'inspirer des articles précédents ou suivants, qu'il voit en totalité sans avoir à ouvrir de fenêtres supplémentaires.
- Il dispose d'un outil qu'il connaît et qu'il maîtrise déjà.

Par contre, cette solution n'est envisageable que dans le cas de dictionnaire relativement simples (dont la structure s'apparente à une structure de traits). C'était le cas pour la plupart des dictionnaires que nous avons été amenés à manipuler .

Par contre, lorsque nous avons voulu informatiser le dictionnaire explicatif et combinatoire du français contemporain (dans le cadre du projet NADIA-DEC), qui présente une structure beaucoup plus complexe, il n'a pas été possible d'utiliser cette technique et nous avons du construire un éditeur spécialisé pour le DEC : DECID.

3. DECID, un éditeur spécialisé pour le DEC

3.1. Motivations

Le projet NADIA-DEC vise la création d'une version informatisée du dictionnaire explicatif et combinatoire du français contemporain. Cette version devra contenir l'ensemble des informations présentes dans le DEC sous une forme aussi structurée que possible. Elle s'appuie donc sur le système de gestion de bases lexicales multilingues SUBLIM défini au GETA-CLIPS ([Sérasset 94]).

Nous ne visons pas d'application particulière des données ainsi informatisées, afin de ne pas privilégier certains aspects de la structure au détriment des autres. Nous souhaitons que le dictionnaire explicatif et combinatoire soit informatisé sans subir de modification fondamentale par rapport a sa version papier. Ceci garantira l'utilisation des outils développés par les rédacteurs de la version papier, ce qui est une condition de la réussite de ce projet (création de la version informatisée avant la version papier). De plus, nous pensons que l'ensemble des informations trouveront une exploitation dans la communauté TALN. Nous souhaitons donc que l'intégralité des informations présentes sur papier soit disponible sous forme informatique à partir de laquelle chacun pourra dériver un dictionnaire spécifique.

Ce projet répond à plusieurs motivations de la part de chacun des partenaires. D'une part, il permet de tester le système SUBLIM en l'utilisant pour un dictionnaire mettant en œuvre des structures complexes. D'autre part, les informations contenues dans le dictionnaire explicatif et combinatoire présentent une richesse que l'on ne trouve dans aucun autre dictionnaire informatisé. Enfin, la mise en œuvre de ce projet suppose la création d'outils informatiques simplifiant la gestion d'un tel dictionnaire.

Pour atteindre ces différents objectifs, nous souhaitons non seulement informatiser une version du DEC, mais surtout informatiser sa chaîne de production afin de faire en sorte que le DEC existe d'abord sous forme informatique puis sous une forme imprimée. Ainsi, le travail à effectuer se décompose en différentes étapes :

1. création d'un éditeur informatique spécialisé pour le DEC,
2. réalisation d'un mécanisme d'import des données existantes, actuellement au format R.T.F. (Rich Text Format),
3. réalisation d'un module d'export vers différents formats utilisables informatiquement (SGML/TEI, format LISP…) et pour la publication papier (R.T.F., M.I.F.…).
4. intégrer dictionnaire et éditeur à un système de gestion de données lexicales générique (SUBLIM). Les vérifications des différentes contraintes sur les données seront vérifiées à ce niveau,

Les point 1, 2 et 3 sont en cours de réalisation et sont assez avancés. Le point 4, à plus longue échéance, en est à ses débuts.

L'éditeur DECID créé au cours du projet sera mis à disposition de l'ensemble de la communauté TALN.

3.2. Originalité

Une autre approche de l'informatisation du DEC a été utilisée, notamment à l'université de Montréal ([Mel'ãuk & al. 1995], chap. 4) par Alain Polguère :

- Pour Alain Polguère, l'informatisation du DEC est un moyen de disposer de données pour des traitements automatiques. Ainsi, seules les données complètement formalisables sont retenues lors de cette informatisation. De plus, certaines de ces données sont légèrement modifiées par rapport à la version imprimée et d'autres sont ajoutées.
- le projet NADIA/DEC vise l'informatisation du processus de production du DEC, il doit donc prendre en compte l'ensemble des informations présentes dans le DEC, et ce, quel que soit leur degré de formalisation.

Les deux approches ne sont pas antinomiques et nous travaillons actuellement à l'adaptation de l'éditeur DECID pour la génération (au moins partielles) d'entrées lexicales utilisables par Alain Polguère. Nous envisageons aussi d'augmenter les fonctionnalités de l'éditeur actuel, afin d'intégrer dans le processus de création d'entrées du DEC, la saisie de données propres au formalisme utilisé à l'université de Montréal.

3.3. Présentation de l'éditeur

Dès que l'on créé ou que l'on ouvre un dictionnaire, la fenêtre principale du dictionnaire apparaît (figure 2). Elle se décompose en deux parties. La première (liste de gauche) contient l'ensemble des vocables définis dans ce dictionnaire. Lorsque l'on sélectionne un (ou plusieurs) vocables dans cette liste, les lexies correspondantes apparaissent dans la liste de droite.

Ces lexies apparaissent sous forme d'un numéro de sens (lorsqu'il y en a) suivi d'un résumé. C'est ce résumé qui est utilisé dans la version papier comme tableau synoptique d'un vocable. Ces résumés sont éditables.

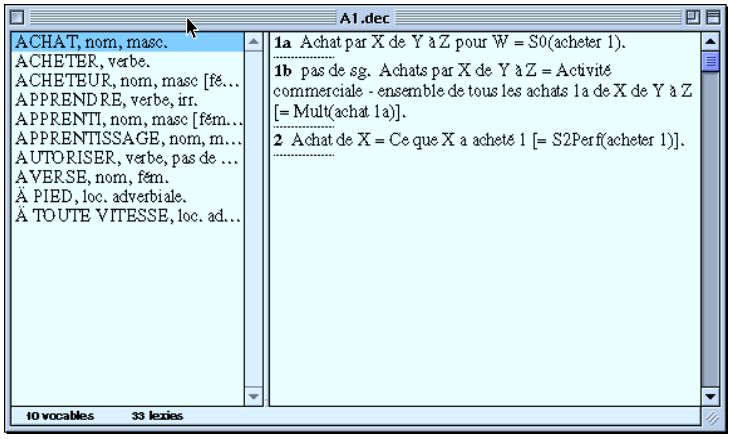

Figure 2 : la fenêtre principale d'un dictionnaire.

En double-cliquant sur un résumé, on ouvre la fenêtre de la lexie correspondante (figure 3).

Il faut noter que l'unité du lexique est la lexie et non le vocable. Les informations représentées dans la colonne de gauche sont calculées « à la volée » à partir de l'ensemble des lexies du dictionnaire. Elles ne sont pas stockées dans le dictionnaire.

La fenêtre de lexie contient les informations graphiques et morphologiques, la définition, les exemples et la liste des fonctions lexicales.

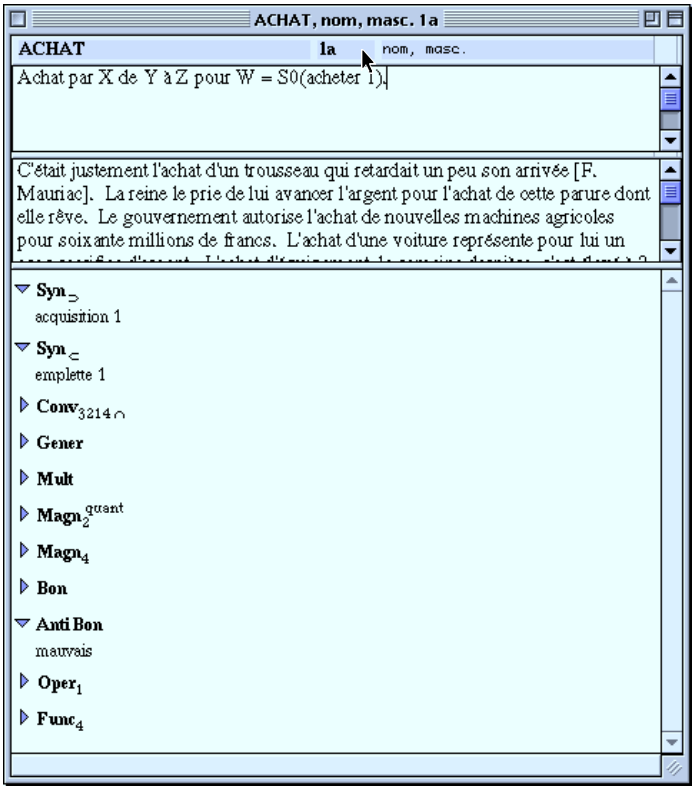

Figure 3 : la fenêtre de lexie.

Dans cette fenêtre, chaque élément est éditable. On peut ainsi à tout moment changer la graphie, la définition, le numéro de lexie ou toute autre information.

Les informations graphiques et morphologiques apparaissent comme dans le dictionnaire papier. Il est possible de spécifier la portée de ces informations (propres à une lexie ou commune à toutes les lexies du vocable).

Les fonctions lexicales sont présentées sous forme de liste. Chaque élément de cette liste est éditable. L'éditeur DECID permet d'éditer de manière très simple les fonctions lexicales. Ainsi, il est possible de créer la fonction lexical composée **LiquOper$_1$** en tapant simplement **l**, **i**, **q**, **u**, **o**, **p**, **e**, **r**, **1**.

Actuellement, l'interface de création des fonctions lexicales ne permet que la création de fonctions standards simples ou composées (figure 4).

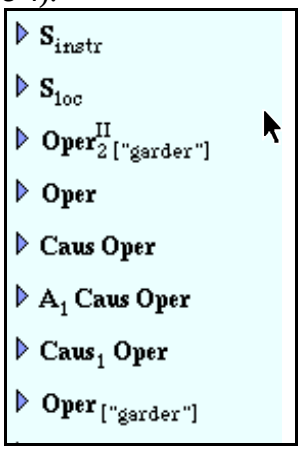

Figure 4 : Fonctions lexicales simples et composées

Elle ne permet pas encore la création de fonctions lexicales non standard ou complexes (figure 5). Par contre, le noyau de représentation d'une fonction lexicale, ainsi que le module d'affichage sont suffisamment avancés pour gérer ce type de fonction. Ainsi, il est possible de récupérer et de visualiser des fonctions lexicales non standard ou complexes à partir des fichier word du dictionnaire publié.

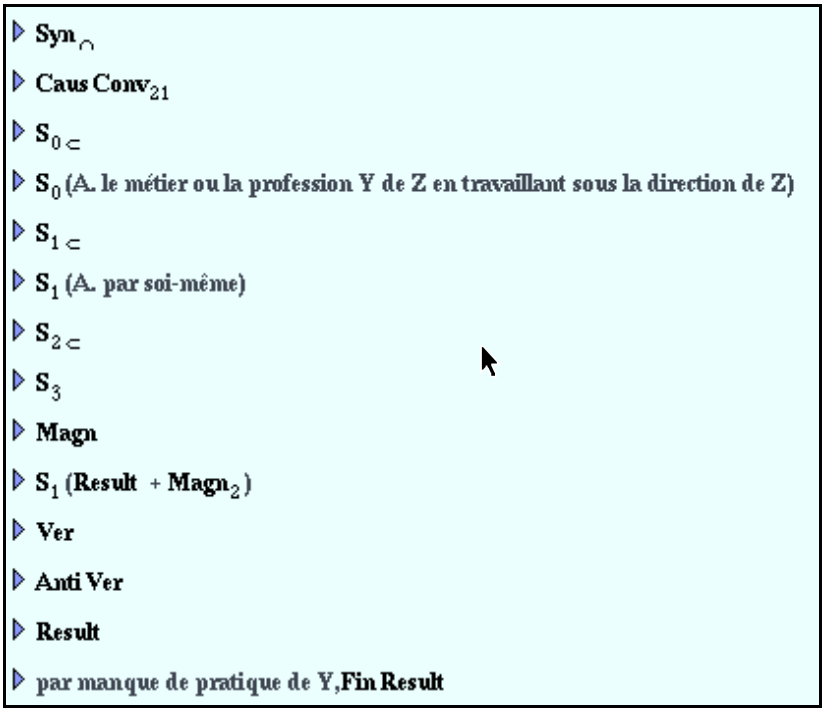

Figure 5 : Fonctions lexicales non standard et complexes

4. Conclusion

Nous avons présenté deux approches différentes de constructions d'outils pour lexicographes. Ces approches sont complémentaires et seront choisies en fonction de la complexité du dictionnaire à indexer.

Ces deux approches répondent différemment au problèmes posés dans la première partie de cet article.

La première permet de distribuer facilement le travail parmi différents lexicographes. De plus, elle est facilement paramétrable puisque les macros word sont générées à partir d'une description des entrées. Un mécanisme de gestion de version nous permet de modifier facilement la structure du dictionnaire et ce, même si des lexicographes sont en train de travailler sur des parties du dictionnaire.

La seconde donne au lexicographe un outil très convivial et offrant de nombreux services spécifiques à la structure particulière du dictionnaire considéré. Par contre, l'éditeur DECID est très dépendant de la structure du dictionnaire. Il est donc très difficile de modifier celle-ci.

Pour remédier à ce problème, nous souhaitons définir, autour de SUBLIM, un outil de création d'interfaces. Ainsi, la modification d'une structure sera possible sans avoir à re-programmer des parties de l'éditeur. Cette approche est possible par l'utilisation d'architectures mettant en œuvre des objets collaborant (comme OpenDoc ou les Java Beans).

Enfin, le travail de lexicographie nécessite l'accès à de nombreuses informations de sources différentes. Ces sources tournant parfois sur des plates-formes informatiques particulières. Nous pensons que l'intégration du travail de lexicographie dans une architecture en réseau est une réponse à cette contrainte. Aussi, l'utilisation du langage Java semble appropriée pour construire une véritable environnement de travail du lexicographe, homogène et portable.

5. Références

Mel'ãuk I., Clas A. & Polguère A. (1995) *Introduction à la lexicologie explicative et combinatoire.* Universités francophones et champs linguistiques, AUPELF-UREF et Duculot, Louvain la Neuve, 256 p.

Sérasset G. (1994) *SUBLIM : un système universel de bases lexicales multilingues et NADIA : sa spécialisation aux bases lexicales interlingues par acceptions,* Thèse nouveau doctorat, Université Joseph Fourier-Grenoble 1 : 194 p.